Quantum chemistry property surface modeling of ^{13}C chemical shifts in a long-lived spin state system

Jari Havisto

University of Oulu
Faculty of Science
NMR Research Unit

26.6.2020

Contents

Abstract

Application of nuclear magnetic resonance (NMR) experiments for nuclei other than proton, such as ^{13}C, can be limited by low signal-to-noise ratio (SNR) (Although, it can be a problem for protons also). Special techniques to increase the SNR can suffer from short signal enhancement lifetimes, usually limited by the spin-lattice relaxation constant T_1. Long-lived spin states (LLS) can increase the polarization lifetimes significantly. LLS in this work is a naphthalene derivative singlet state that has a pair of ^{13}C spins in it's center, with a small chemical shift difference (dCS). A small dCS is required to access the spin-state with NMR-methods. This small dCS can be difficult to find experimentally so the aim of this work is to get an idea if a change of temperature increases this difference. In principle, the ensemble averages of the dCSs could be computed with quantum chemistry (QC) methods. However, computing enough accurate dCS values with these methods is impractical due to large computational load. Also, a large number of calculations would be needed to get a meaningful estimate of the dCS ensemble average even for one temperature. Therefore, the computational load is largely circumvented here by building a quantum-chemistry property surface (QCPS) model. The model uses QC-computed shielding constant values and interpolates them as a function of spatial degrees of freedom (DOF) of the molecule. This can possibly allow for a meaningful reduction in the number of required QC computations. The model can then be used to analyze on molecular dynamics (MD) trajectories simulated at different temperatures and a change of the ensemble average of dCS can be determined. QC computations are a bottleneck that limits the QCPS model dimensions. Therefore, this work also investigates which degrees of freedom the model should contain. Also a second QCPS model is optimized with smaller basis set of QC computed training set in order to compare the effect of a less accurate training sample to the model's performance. The resulting ddCS/dT-slope is dominated by statistical error due to model's low dimension of 28 DOF. This model lacks predictive range and the obtained dCS correlation coefficient with explicitly QC-computed test set of 0.7049 is low. The correlation coefficient of the model optimized with less accurate training samples is 0.7050, so both models have similar correlations with their corresponding test sets. In order to increase the model performance, an investigation on the required model dimension is suggested with, possible addition of solvent molecules.

Jari Havisto

1 Introduction

Nuclear magnetic resonance (NMR) has applications in many scientific fields. A NMR experiment is, however, fundamentally limited by the properties of the sample. Typically receiving a good signal-to-noise ratio (SNR) from other than proton nuclei is challenging and special techniques are routinely applied for the signal enhancement. For example, the ^{13}C spectrum of organic molecules can be enhanced by hyperpolarization techniques, such as dynamic nuclear polarization (DNP) [1]. This signal enhancement has many applications, one being in the characterization of human cancer [2]. However, lifetimes of these signal enhancements are relatively short. Typically they are in the order of tens of seconds [3] and are usually limited by the spin-lattice relaxation constant T_1 [4, 5]. This relatively short lifetime of the polarization limits experiments such as slow diffusion, flow and metabolic conversion [6]. By converting the magnetization into long-lived state (LLS) these lifetimes can be significantly increased [7]. In this work, the LLS is a singlet state of a naphthalene derivative (see Figure 1 and Sections 1.1, 2.3). Important aspect of this system is a pair of $\frac{1}{2}$−spins in nearly equivalent chemical environments, which are relatively well isolated from other spin-active nuclei. This isolation guards the spins from the fluctuations of the intramolecular dipole-dipole couplings induced by molecular motion. This is an important mechanism for spin-lattice relaxation [8]. This LLS can allow polarization lifetimes T_S that are much longer than the typical lifetimes determined by T_1. It has been demonstrated that singlet lifetimes over one hour in room temperature are possible [8].

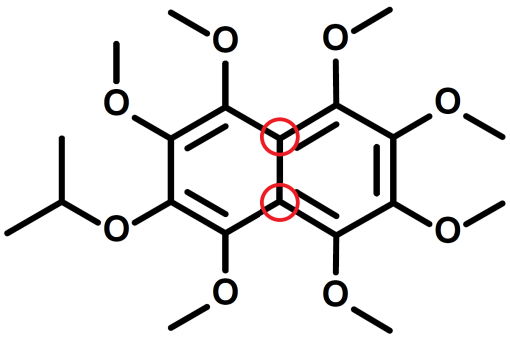


Figure 1: Molecule 1,2,3,4,5,6,8-heptakis([d_3]methoxy)-7-(([d_7]propan-2-yl)oxy)-naphthalene sketched with Chemdraw. The two red circles identify the two carbon-13 nuclei of interest.

Singlet-state lifetime T_s is dependent on the isotropic chemical shift difference (dCS) between the shielded spin-pair, $\Delta\delta_{iso}$ [9]. The aim of this work is to construct a quantum chemistry property surface (QCPS) model, which allows for an efficient prediction of the mean dCS of a simulated molecular dynamics (MD) trajectory at a given temperature. This approach could then be used as theoretical prediction of the direction

of change in the dCS as a function of temperature. Although a small isotropic shielding constant difference minimizes the relaxation caused by singlet-triplet leakage [9], a small difference is needed in order to access the singlet-state with NMR-pulse methods [5]. However, this small difference can be difficult to find experimentally. Therefore, an understanding on whether the dCS can be manipulated by changing the temperature can be a great benefit when experiments are designed.

For this work, the behavior of the molecule in a given temperature has been simulated with MD software Amber [10]. Snapshots of this simulation can be exported as MD trajectory, from which an ensemble average of the chemical shift difference can be computed. Calculating enough accurate $\Delta\delta_{iso}$ values for a meaningful ensemble average by "brute force" quantum chemistry (QC) methods alone is not feasible, even when the computational load is lightened by means such as the locally dense basis-set (LDBS) approach (see Section 2.6). Therefore, the above-mentioned QCPS model is sought here (see Section 2.8). This type of model has previously been used successfully in [11] to estimate weak interaction that are important in singlet-state relaxation. It has also been used for proton chemical shift estimation in [12] (referred as Gaussian process). The model takes quantum-chemically calculated tensor values as ground truth points and interpolates them as a function of spatial degrees of freedorm (DOF) (see Section 3.4). Output of the model is considered to be an output of a random multivariate gaussian process, where mean is the predicted value and standard deviation can be taken as a measure of error in a given point. This can be used to identify where the model needs to be improved the most and congruent sampling, i.e, which samples of the MD trajectory are most beneficial to include, can be done. This work uses QC-computed shielding constants of the configurations picked from the MD trajectories as model ground truth values.

Efficient sampling of the whole MD trajectory ($\mathcal{O}(10^5)$ snapshots) is paramount for meaningful results. $\mathcal{O}(10^2)$ training samples (and test samples) are selected by maximizing Euclidean distances between degrees of freedom. This is a simple way that was utilized in this work to ensure that the selected configurations are in a sense "maximally separated" and by large, a good representation of the limits of the MD trajectory. Using the selected training set, interpolation model can be optimized using selected DOFs. The model can then be used to infer shielding tensor values for the whole MD trajectory. If the predictions are accurate, this should yield a meaningful ensemble average for the chemical shift difference. If this model can then be used for another trajectory, that is simulated at different temperature, an estimate of $\mathrm{d}dCS/\mathrm{d}T$-slope can be found, which gives information on how the dCS roughly behaves with the change of temperature. This result can then be verified with existing experimental measurement.

However, the first step in this work is to understand the spatial DOFs of the

molecule and their relative importance for the dCS. Because the QC computations produce a significant bottleneck for the training set points, the sought QCPS model is limited in its dimensions. Therefore, a set of small models is built as a first step in order to understand the relative contributions of the DOF-groups (see Section 3.4 and 4.1). This step should allow the selection of best DOFs to produce the best correlation with the QC-computed dCS test set values.

The accuracy of the QCPS-model predictions are also dependent on the accuracy of the QC-computed isotropic shielding constant differences. The model can take into account the uncertainties for these values [11] but here, the training values are assumed to be absolutely accurate for simplicity. To investigate the required level of accuracy in QC computations, two training sets (and test sets) are computed with two levels of basis sets (see Methods Section) using density-functional theory (DFT).

In the next Section, this thesis first introduces the naphthalene derivative molecule under investigation. After that, previous work done in minimizing the QC bottleneck is discussed. Section 2 describes the relevant theory, followed by methods, results and discussion. Appendix describes how the calculated shielding constants are interpreted as chemical shift differences. The steps in this work are summarized visually in Figure 3 as a flow chart.

1.1 Naphthalene derivative molecule

The molecule investigated in this work is the naphthalene derivative: 1,2,3,4,5,6,8-heptakis([d_3]methoxy)-7-(([d_7]propan-2-yl)oxy)-naphthalene (Figure 1). The two middle carbons are isotope-13, referred here as the spin pair. Important to note is that the molecule is nearly inversion symmetric, only having one differing side group connected to the carbon rings. This causes a small isotropic chemical shift difference that is required to convert the triplet-state to singlet-state and vice versa [13]. No solvent molecules are included in this study.

1.2 Decreasing computational load

QC computations for the naphatlene derivative are relatively costly. Therefore, the computational load of the training set needs to be considered closely. Isotropic shielding constants for the training set are computed using DFT. The computational load of DFT calculation is by large determined by the number of basis functions used to approximate the molecular orbitals (see Section 2.5). Related previous work [14] investigated systematically the accuracy vs. computational cost of basis sets in the naphthalene derivative. The report utilized aug-pcSseg-n/pcSseg-n family basis sets and the (LDBS) approach [15], where larger basis sets are used for different parts of the molecule. This method saves basis functions from the "less important" parts of

the molecule, keeping the computational load lighter while ideally maintaining a high accuracy. Ideal balance between accuracy and computational cost was achieved when aug-pcSseg-3 basis set was used for the middle carbon-13 isotopes and aug-pcSseg-1 for the rest of the molecule. The report also found that the calculations were insensitive for the change of certain numerical parameters. This allows for the use of "coarser" computations that are much faster to converge in the self-consistent field iterations. These results are used as the baseline when computing the training set tensors.

2 Theory

This section is for the most part adopted from the author's previous work [14].

2.1 Nuclear magnetic resonance

NMR results from an interaction between the magnetic moments of nuclei and the external magnetic field. This effect is studied on basic nuclear physics courses. The interest here is not the mathematical expressions, but the basic understanding of the phenomenon and the relevant terms concerning this report. The connection between chemical shift (CS) and the shielding constant is expanded in Appendix A. What is represented here is a basic energy-state model that explains simple NMR experiments. This model works well enough for the purposes of this report. In order to understand NMR more accurately, more detailed quantum-mechanical approach would be needed. More about methods and theory of NMR can be found for example in Ref. [4].

Nuclei that have non-zero magnetic moment can be detected with NMR methods. In the presence of a strong external magnetic field, the different spin states are separated into different energies. The energy difference between these states corresponds to a photon of frequency ν, called the Larmor frequency. In thermal equilibrium, there is a small population difference between these states that causes a small net magnetization. In order to get a measurable signal, spins from lower spin-state have to be excited to a higher spins-state. This is done with radio frequency pulse sequencing. When this excitation relaxes, photons are emitted and detected as the NMR signal. [16]

2.2 Chemical shift and shielding

Electrons can affect the local magnetic field that a nucleus experiences. The external magnetic field B_{ext} induces electron orbital angular momentum which produces a small magnetic field $\vec{B}_{ind}$ (11) at nucleus i. The local magnetic field $\vec{B}_{local}$ becomes

$$\vec{B}_{local} = \vec{B}_{ext} + \vec{B}_{ind} = (\mathbf{1} - \boldsymbol{\sigma}_i) \cdot \vec{B}_{ext}. \tag{1}$$

Here, the local field is expressed with the help of the shielding tensor $\boldsymbol{\sigma}_i$. However, this report is only interested in the isotropic shielding constant σ_i: the average of the diagonal elements of $\boldsymbol{\sigma}_i$. Now because the local field is changed, the Larmor frequency is changed to

$$\nu_i = (1 - \sigma_i)\frac{\gamma_i B_{ext}}{2\pi}, \tag{2}$$

where γ_i is the nucleus-specific gyromagnetic ratio.

This "shift" in the resonance frequency is called the CS and it gives information about the local chemical environment of the nucleus. In the naphthalene derivative of present interest, one side group is different from the other side group (Figure 1). Most of the groups connected to oxygen are CD_3 but one group is $CD(CD_3)_2$. This produces a small dCS in the central carbon-13 pair, indicated with red circles in Figure 1. The CS of Larmor frequency is dependent on the external magnetic field being used. Therefore, it is customary to express CS in a scale that removes this dependency: the delta scale expressed in ppm (parts per million),

$$\delta_i = \frac{\nu_i - \nu_{ref}}{\nu_{ref}} \cdot 10^6. \tag{3}$$

ν_{ref} and ν_i are the resonant frequencies for the reference standard and for the nucleus i respectively. δ_i is the CS expressed in ppm for the nucleus i. Difference between the shielding constants is defined as $\Delta\sigma = \sigma_3 - \sigma_2$ (see Figure 4). For carbon-13, the standard often used is the carbon-13 frequency of Tetramethylsilane $Si(CH_3)_4$ (TMS). The relationship between the dCS and the difference between the shielding constants is considered to be

$$\Delta\delta \approx -\Delta\sigma \tag{4}$$

in this report, which is derived in Appendix A. [16]

2.3 Criteria for long-lived spin state

In this work, the LLS is a singlet state, meaning that the nuclear spin wave function is anti-symmetric for the exchange of the two atoms. In the naphthalene derivative, there are two spin-$\frac{1}{2}$ nuclei of interest: the carbon-13 isotopes at the center of the molecule (Figure 1). The spin pair is in relative isolation from other spin-active nuclei in order to minimize relaxation caused by dipolar or scalar coupling with them (more about coupling can be found for example in Ref. [16]). Nuclei of the spin pair have a slight difference in their local magnetic environments, caused by one side group of the molecule being different. This has to be so in order to the spin state to be accessible by NMR methods. Other relaxation effects are not considered in this report, more about them can be found in ref. [8].

2.4 Born-Oppenheimer approximation

In order to calculate the shielding constants for the atoms of interest, the Schrödinger equation (SE) needs to be solved. This cannot be done analytically even for the simplest molecule $H_2{}^+$, so approximations are necessary. A good starting point for solving the SE for molecules is the Born-Oppenheimer approximation. The idea behind it is the huge difference in the mass of nucleus and electrons. Light electrons can be thought of as moving in a stationary potential generated by the heavy nuclei in the molecule. Now the wave function can be separated into two parts

$$\Psi(\vec{r};\vec{R}) = \psi(\vec{r};\vec{R})\chi(\vec{R}), \tag{5}$$

where ψ is the electronic wave function and χ is the nuclear wave function. In the convention adopted here, $\psi(\vec{r};\vec{R})$ means that the function is dependent on the position $\vec{r}$ and parametrically dependent on the position of the stationary nuclei $\vec{R}$. By the separation of variables method, the SE can be solved for the electronic wave function and nuclear wave function separately. [17]

2.5 Density-functional theory

In DFT, the electronic energy can be expressed in terms of $\rho(\vec{r})$, which is the total electron density at the point $\vec{r}$. Electronic energy is a functional of electron density. In other words: electron energy is a function of electronic density $\rho(\vec{r})$, which is a function of position $\vec{r}$.

$$\begin{aligned} E[\rho] = -\frac{\hbar^2}{2m_e}\sum_{i=1}^{n}\int \psi_i^*(\vec{r_1})\nabla_1^2\psi_i(\vec{r_1})d\vec{r_1} - j_0\sum_{I=1}^{N}\frac{Z_I}{r_{I1}}\rho(\vec{r_1})d\vec{r_1} + \\ \frac{1}{2}j_0\int\frac{\rho(\vec{r_1})\rho(\vec{r_2})}{r_{12}}d\vec{r_1}d\vec{r_2} + E_{XC}[\rho]. \end{aligned} \tag{6}$$

Here $E_{XC}[\rho]$ is the exchange-correlation energy, $\hbar$ is Planck's constant divided by 2π, m_e is the mass of an electron, n and N are the number of electrons and nuclei, respectively. ψ_i are Kohn-Sham (KS) orbitals, $j_0 = \frac{e^2}{4\pi\epsilon_0}$ and ρ is the electron density. Because the form of the exchange-correlation functional is unknown, it has to be approximated. The exchange-correlation functional used in the calculations of this report is KT2 [18] (see Section 3.3). As a first step in DFT, an initial guess of the electron density is made. Then, using the chosen exchange-correlation functional, exchange-correlation potential

$$V_{XC}[\rho] = \frac{\delta E_{XC}[\rho]}{\delta\rho} \tag{7}$$

is computed. Next, from the KS equations

$$\left\{-\frac{\hbar^2}{2m_e}\nabla_1^2 - j_0\sum_{I=1}^{N}\frac{Z_I}{r_{I1}} + j_0\int\frac{\rho(\vec{r_2})}{r_{12}}d\vec{r_2} + V_{XC}(\vec{r_1})\right\}\psi_i(\vec{r_1}) = \epsilon_i\psi_i(\vec{r_1}), \tag{8}$$

the initial set of KS orbitals ψ_i are solved. The KS orbitals can be solved with basis set expansion using basis functions θ_j

$$\psi_i = \sum_{j=1}^{M} c_{ji}\theta_j, \tag{9}$$

where c_{ji} are unknown coefficients and M is the number of basis functions. The basis sets used in this basis set expansion are the focus of the previous work [14]. This method reduces the solving of the KS orbitals to determination of coefficients by linear system of equations: matrix manipulations. Finally, from the set of orbitals, an improved density

$$\rho(\vec{r}) = \sum_{i=1}^{n}|\psi_i(\vec{r})|^2 \tag{10}$$

is calculated. This process is repeated until the results are self-consistent, meaning that that following iterations change the result less than a pre-selected small amount. In this self-consistent field (SCF) approach, the energy is minimized while the total charge remains constant. Lastly, the electronic energy is calculated from equation (6). For more information about the topic, the reader is directed to relevant literature. [17]

2.6 Basis sets and the LDBS approach

In the previous Section, the KS orbitals were solved using a basis set expansion (9). This means that the orbitals are expressed as linear combinations of some type of basis functions. This approach is exact only if the basis set expansion is complete and the number of basis functions is infinite. This is, however, impossible, so the expansion necessarily needs to be finite. In general, the higher the number of basis functions, the better is the representation of a single orbital. In principle, many types of functions can be used. However, a good choice of function type in the expansion requires a small number of basis functions to achieve required accuracy. The computational cost increases rapidly for DFT and Hartree-Fock (HF) (at least as $\mathcal{O}(M^3)$) [19], so the increase in the number of basis functions significantly increases the computation time.

A good choice to represent atomic orbitals are the Slater-type orbitals (STO), since these mirror the solutions to SE of the hydrogenic atom. They are, however, computationally expensive. They can be approximated with Gaussian-type orbitals (GTO), that are computationally inexpensive, but are individually poor representations of

atomic orbitals. One issue is that, at the nucleus, the derivative of STO's is discontinuous, while a GTO has a zero derivative. Also, the tail end of a primitive GTO decreases more rapidly than of a STO which has the correct long-distance functional form. These issues can be overcome by using many GTOs to approximate one STO. Even though this increases the number of basis functions, the computational ease of GTO makes this worthwhile.

In a minimal basis expansion, only one basis function is used for each orbital. The notation STO-nG means that n primitive gaussians are used to describe one STO. For example, STO-3G notation would mean, that for carbon atom ($1s^2 2s^2 2p^2$), altogether six primitive GTOs are used to represent the two STO-type s-orbitals, and 3 GTOs for each STO-type $p_{x,y,z}$-orbital. The minimal expansion is, however, not able to describe molecular bonding and charge polarization very well. The basis set can be improved by adding additional polarization functions (denoted by * after basis-set name in Pople-style basis sets [20] and with ** if such functions are added also for the hydrogen atoms). These are one step higher angular momentum atomic orbitals than the largest angular momentum orbital in the ground-state atom. These functions can improve the flexibility of description. For hydrogen, this would correspond to addition of a single p-type orbital. These functions also serve in describing electron correlation effects.

In order to add more flexibility to the description of molecular bonding, the number of functions used to describe one orbital can be doubled (double-Zeta DZ), tripled (triple-Zeta TZ) etc. For example, a double-zeta basis set for carbon would be (1s, 1s', 2s, 2s', $2p_x$, $2p_y$, $2p_z$, $2p'_x$, $2p'_y$, $2p'_z$) where each function can again be represented with n primitive GTOs.

When basis sets are energy-optimized (GTOs optimized to reproduce minimum energies in the systems used to train the basis set), the energy dependence is mainly in the inner-shell electrons. However, in chemistry all the bonding happens with outer-shell, valence electrons. To overcome this, basis sets are sometimes enhanced with inclusion of small exponent basis functions: diffuse functions (augmented basis sets, often denoted by aug). These are important in properties that depend on tail part of the wave function, such as electric moments and polarizabilities.

A common basis sets in use are the Pople-style basis sets [20], noted as k-nlmG. k notes how many primitive GTOs are used to describe core orbitals. n,l and m denote how many primitive GTO's are used for valence orbitals. Possible polarization functions are noted after G. For example 3-21G would be double-zeta quality in the valence, with inner shells formed by contracting 3 primitive Gaussians. Two valence orbitals are contracted from 2 and 1 Gaussians, respectively.

The basis-set types of most interest to this report are so called polarization-consistent basis sets, pc-n [21]. These are designed for fast SCF convergence in DFT calculations. Integer n denotes the level of polarization beyond atomic system; n=0 is unpolar-

ized, n=1 has one polarization function (double-zeta quality) and n=2 is triple-zeta quality etc. For NMR properties calculations the addition of single "tight" p-function with large exponents (as in close to nucleus) to pc-n basis sets yields the pcS-n basis sets [22]. This family of basis sets are optimized for NMR calculations (S stands for shielding).

Primitive gaussians can be contracted, meaning that some of the basis functions are used in a fixed linear combination. This allows the number of basis functions to be reduced. This is especially effective in core electrons, since they are largely independent of the environment. This will, however, reduce accuracy, because it will result in lesser flexibility. In general contraction, the contracted GTOs are constructed using all primitive Gaussians. In the segmented approach, one primitive is allowed to be used only in one contracted Gaussian. This contraction is noted as "seg" in the basis set name. This leads to a family: pcSseg-n [23], which is further optimized for DFT shielding calculations, and is the type used in this report. [24]

For larger molecules, the computational load is a significant bottleneck for large number of calculations. One effective way to reduce the number of basis functions in QC calculations is to use the LDBS approach [15]. This method attempts to focus basis functions on certain spatial regions of the molecule. This can be achieved simply by using larger basis sets for the parts of the molecule that are assumed to be the most important in a given computation. This approach is fundamentally a balancing act in the desired accuracy and the computational cost.

2.7 NMR property calculation

Defining the unperturbed reference states with the approximately determined wave functions and energies, a number of different property calculations can now be performed with perturbation theory. The shielding constant is defined as

$$\vec{B}_{ind}(\vec{r}) = -\sigma(\vec{r})\vec{B}_{ext}. \tag{11}$$

First step to calculate the induced magnetic field $\vec{B}_{ind}$ is the Biot-Savart law, classically

$$\vec{B}_{ind}(\vec{s}) = \frac{\mu_0}{4\pi}\int \frac{\vec{j}(\vec{r}) \times (\vec{s}-\vec{r})d\tau}{|\vec{s}-\vec{r}|^3}, \tag{12}$$

where μ_0 is the vacuum permeability. It can be used to calculate magnetic field induced by known current density distribution. This can be used as a stepping stone for response theory approach in quantum mechanics, which will not be detailed here. In our case, the current densities $\vec{j}$ are calculated from previously determined wave functions, KS orbitals, quantum mechanically in the presence of external magnetic field. The field-dependent hamiltonian can be formed by replacing the momentum operator $\hat{p}$ with

operator $\hat{p} + \vec{A}_{ext}$. Here, $\vec{A}_{ext}$ is the vector potential of the external magnetic field. Vector potential can be chosen to be divergence-free and its curl is the external magnetic field

$$\nabla \cdot \vec{A}_{ext} = 0 \tag{13}$$

$$\nabla \times \vec{A}_{ext} = \vec{B}_{ext} \tag{14}$$

This still leaves many different choices for $\vec{A_{ext}}$. One possible choice is

$$\vec{A}_{ext} = \frac{1}{2}\vec{B}_{ext} \times (\vec{r} - \vec{R}) \tag{15}$$

where $\vec{R}$ is an arbitrary choice of reference called the gauge origin. The vector field $\vec{A}_{ext}$ is dependent on the choice of gauge origins, however the current density $\vec{j}$ is not. The problem of gauge origin is solved with a method of gauge-including atomic orbitals (GIAO), in which individual gauge origins are assigned to the center of atomic orbitals. More about this can be found in Ref. [25].

2.8 QCPS model

The QCPS model is a Kriging interpolation model, aka. Gaussian process or DACE (Design and Analysis of Computer Experiment) [11]. Interpolation is done globally, meaning that the interpolation is done taking all the training points into account. The model is also capable of estimating errors locally. Here, the training set vector is denoted as $\mathbf{\Theta} = [\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(K)}]^T$, where $\Theta^{(i)}$ denotes all the n degrees of freedom of the ith MD snapshot. The QC-computed tensor components are denoted as $C_{kl}(\Theta^{(i)})$. Starting point for the model is the assumption that the training set values are samples from a K-dimensional probability distribution

$$\boldsymbol{C}_{kl}(\mathbf{\Theta}) = \beta\mathbf{1} + \boldsymbol{Z}(\mathbf{\Theta}), \tag{16}$$

where β is usually a constant (can also be an interpolation polynomial), $\mathbf{1}$ is a unit vector and $\boldsymbol{Z}(\mathbf{\Theta})$ is Gaussian stochastic function with zero mean and covariance of

$$\mathrm{Cov}(\boldsymbol{Z}(\mathbf{\Theta}, \mathbf{\Theta'})) = \sigma_Z^2 \boldsymbol{R}(\mathbf{\Theta}, \mathbf{\Theta'}). \tag{17}$$

σ_Z^2 is called process variance and $\boldsymbol{R}(\mathbf{\Theta}, \mathbf{\Theta'})$ is correlation matrix, defined as

$$\boldsymbol{R}(\Theta^{(i)}, \Theta^{(j)}) = \exp[\sum_{l=1}^{n} \rho_l |\Theta^{(i)} - \Theta^{(j)}|^2]. \tag{18}$$

The model has two global parameters; β and σ_Z^2 and one set of local parameters ρ_l that are specific to a given degree of freedom Θ_l. These parameters are determined with maximum likelihood estimation (MLE) by minimizing negative log-likelihood function.

This yields following estimates for β and σ_Z^2:

$$\hat{\beta} = \frac{\mathbf{1}^T \boldsymbol{R}^{-1} \boldsymbol{C}_{kl}}{\mathbf{1}^T \boldsymbol{R}^{-1} \mathbf{1}} \tag{19}$$

and

$$\hat{\sigma_Z^2} = \frac{1}{K} (\boldsymbol{C}_{kl} - \mathbf{1}))^T \boldsymbol{R}^{-1} (\boldsymbol{C}_{kl} - \mathbf{1}). \tag{20}$$

Differentiating log-likelihood function with respect to ρ_l does not yield an analytical solution. Therefore, an iterative optimization process needs to be deployed. For this purposes, the DIRECT algorithm [26] was used (see next section).

The output of the model C_{kl} for an input $\hat{\Theta}$ is given by

$$C_{kl}(\hat{\Theta}) = \beta + [R(\hat{\Theta}, \Theta^{(1)}), R(\hat{\Theta}, \Theta^{(2)}), \ldots, R(\hat{\Theta}, \Theta^{(K)})] \boldsymbol{R}^{-1} (\boldsymbol{y} - \mathbf{1}\beta) \tag{21}$$

where $\boldsymbol{y}$ is a training set of computed QC tensors values. It is important to note that the training set is explicitly included in the model. If a training set point $\Theta^{(i)}$ is inserted in the model, output equals the corresponding QC-computed tensor $C_{kl}(\Theta^{(i)})$. [11]

In this thesis, the training-set tensor values are the individual chemical shifts of the middle carbon-13 nuclei. As an output, the model yields two-dimensional $\hat{\beta}$ and $C_{kl}(\hat{\Theta})$ values (separate values for both middle carbon-13 nuclei) and the correlation matrix $\mathbf{R}$ is expanded for two dimensions [27].

2.9 DIRECT optimization algorithm

DIRECT (DIviding RECTangles) [26] is a direct search technique that is employed here to find the global minimum of the multivariate function, *i.e.*, optimizing ρ_l parameters in MLE.

Since the degrees of freedom are normalized between $[0, 1]$ the search space is n-dimensional unit hypercube. As the algorithm proceeds, this space will be partitioned into hyperrectangles, each with a sample point in its center. The first step is to evaluate function at center c_1 and set the $f_{min} = f(c_1)$ (step 1. in Figure 2). Next step is to evaluate the set of potentially optimal rectangles by

$$f(c_j) - Kd_j <= f(c_i) - Kd_i, \quad \text{for all } i = 1, 2, 3, \ldots, m, \tag{22}$$

$$f(c_j) - Kd_j <= f_{min} - \epsilon |f_{min}|. \tag{23}$$

$K > 0$ and m is the number of hyperrectangle centers to be sampled and d_i is the distance of the vertices from the central point c_i. Now every rectangle in the set of potentially optimal rectangles is evaluated. Starting from the center c_i, hyperrectangles

are evaluated at the points $c \pm \delta \boldsymbol{e_i}$, where $\boldsymbol{e_i}$ is the unit vector of dimension i and δ is one-third of the maximal side length of the rectangle. This is because the rectangles are divided into thirds and divisions are done only along the long dimensions to ensure shrinking along every dimension. The order of which dimensions are divided first matters, so the following rule is adopted: The dimension where the function has the smallest value (function is evaluated only at center of rectangles) is evaluated first $w_i = \min\{f(c + \delta \boldsymbol{e_i}), f(c - \delta \boldsymbol{e_i})\}$ (step 2. in Figure 2). Then the next smallest dimension and so on until all the dimensions with the maximal side length are divided into thirds. f_{min} is updated and the process is repeated until all the potential rectangles are evaluated. A set number of iterations is carried out (step 3 to N in Figure 2). For more information, see the relevant reference.

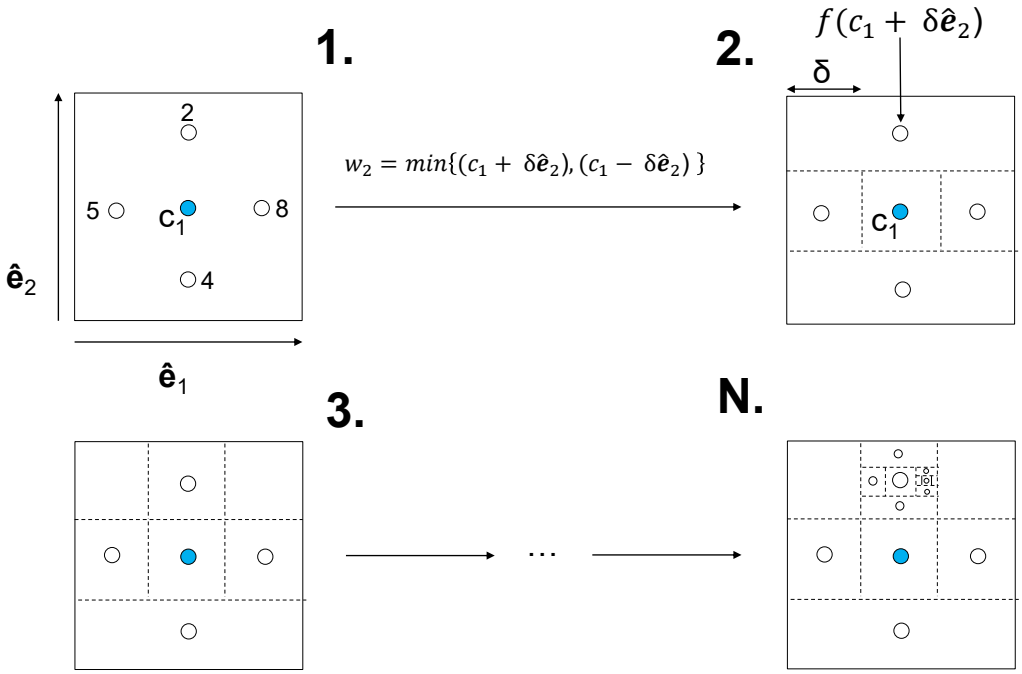


Figure 2: Two dimensional illustration of the DIRECT optimization algorithm. **1**) Center of the search space is picked as a starting point. **2**) The rectangle is first split into thirds along w_2, then in the other direction *etc.*, until rectangle is split along all dimensions. **3**) smallest point is now taken to be the new c_i and process is repeated. When splitting rectangles, only long dimensions are considered. **N**) Process is repeated for a set number of iterations. Illustration inspired by [26].

2.10 Monte Carlo error

A rough way to estimate the error ranges in the mean dCS predictions is to use the Monte Carlo error [28]. The MD trajectory is divided in N equal sections and an ensemble average is computed for each section. The standard deviation σ between the

sections can be considered as the standard deviation of a given measurement. Monte Carlo error is given by

$$\sigma_M \approx \frac{\sigma}{\sqrt{N}}. \tag{24}$$

3 Methods

3.1 The big picture

Figure 3 outlines the process used in this work. Based on the previous work [14] a good LDBS approximation (see Section 2.6) for the molecule is to use aug-pcSseg-3 for the middle carbons and aug-pcSseg-1 for the rest of the molecule (aug-pcSseg-3/aug-pcSseg-1). MD simulation trajectories for 25°C and 35°C were generated with AMBER program (Assisted Model Building with Energy Refinement) [10]. 700 outputs were selected for training and testing (see Section 3.2 and Figure 3 step 1). Using these samples, two training/testing sets were calculated with QC-methods (see Section 3.3 and Figure 3 step 2). The dCS-values of the geometry-optimized structure were also calculated using these basis sets. The computationally more demanding training set (**Set 2**) was then used to investigate and select the best DOFs to include in the full 28-DOF model (Figure 3 step 3). Using the best DOFs, the two QC-computed training sets (**Set 1** and **Set 2**) were then used to construct two QCPS models: **Model 1** and **Model 2** (Figure 3 step 4). These models were then used to calculate all the points in both MD trajectories (25°C and 35°C) and the ensemble average was computed as a mean of a given trajectory. The error ranges of the ensemble averages were estimated with the Monte Carlo error method and results were compared to experimental data (Section 3.5).

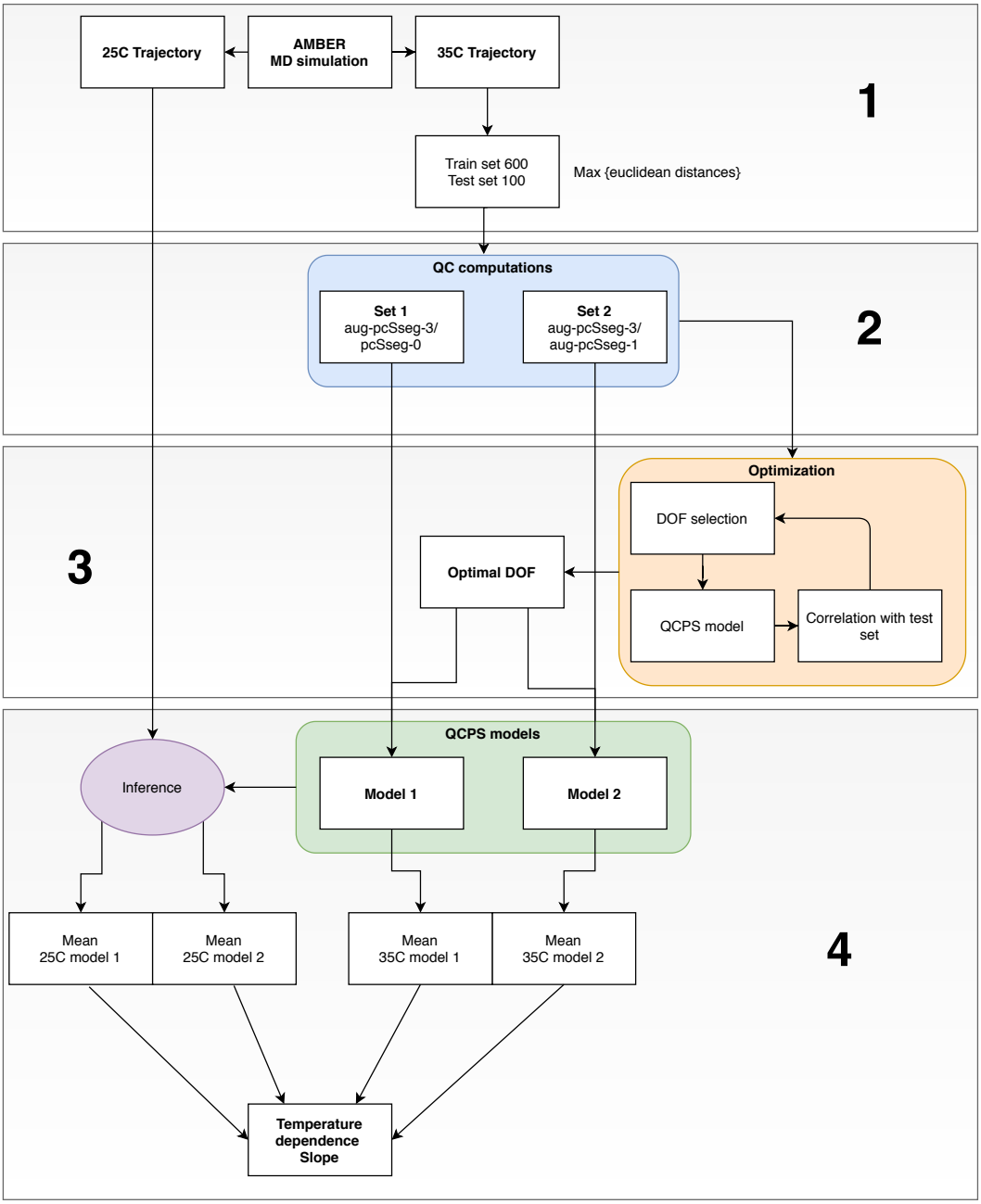


Figure 3: Flow chart shows the different steps taken to achieve the results in this work. **Major steps** are indicated with bold numbers (**1-4**). **1)** The **MD trajectories**, at 35°C and 25°C, were simulated with AMBER-program and 600 training and 100 test samples were selected using the euclidean distance maximization (see Section 3.2). **2)** Two training/test sets were computed with two QC models: aug-pcSseg-3/pcSseg-0 (**Set 1**) and aug-pcSseg-3/aug-pcSseg-1 (**Set 2**). **3)** Using the QC-computed Set 2, different DOF-groups were investigated heuristically to determine an **optimal DOF**-selection for the QCPS models. **4)** Two QCPS models are build with the optimal DOF and optimized (via machine learning) with the corresponding training sets, thus generating two QCPS models: **Model 1** and **Model 2**. The optimized model can then be used to infer the dCS values from the entire MD trajectory, resulting in a mean dCS value for a given model. The 35°C-optimized model can also be used to analyze the 25°C-trajectory, resulting in a **ddCS/dT-slope**.

3.2 Molecular dynamics simulation and sampling

The two MD trajectories (25°C and 35°C) were simulated with AMBER-program using the Carpo cluster-computer (see Figure 3 step 1). The vacuum structure was optimized using DFT with B3PW91/6-31G(d,p) [8]. The particle mesh Ewald method was used for the electrostatic interactions with a cutoff at 11 Å. The equilibration was done over 1.2 ns with 1 bar as the target pressure NTP. The productive runs were done over 1.8 ns for the 35°C trajectory (36000 snapshots) and 1 ns for the 25°C trajectory (25000 snapshots), using velocity Verlet time step integration algorithm with 1 fs time step NVT with weak coupling to a thermal bath. [10]

From the MD trajectory of $\mathcal{O}(10^5)$ snapshots, a training set of 600 and a test set of 100 samples were selected. This was done by selecting 700 samples using

$$\min_{\Theta(K+1)\in\{\Theta\}Np} \sum_{i=1}^{K} \sum_{j=i+1}^{K+1} \frac{1}{\|\Theta(ij)\|^2} \tag{25}$$

where $\left\|\Theta(ij)\right\|^2$ is the euclidean distance between the parameters of the configurations i and j [11]. This minimization finds the MD snapshot that has the largest Euclidean distance compared to previous selections. The last 100 selected snapshots were taken as the test set. Training set is the set of configurations that were used to build the model. Test set is the set used to test the model's performance.

3.3 Quantum-chemical calculations

QC tensors for the training set, *i.e.*, shielding constants σ were calculated using the Dalton 2016 program [29] (see Figure 3 step 2). Calculations were carried out in the FGCI (Finnish Grid and Cloud Infrastructure) grid [30] with KT2 as the Exchange-correlation functional. Two training/test sets of 700 samples each were calculated using aug-pcSseg-3/pcSseg-0 (Set 1) and aug-pcSseg-3/aug-pcSseg-1 (Set 2) basis sets. Shielding constants of the middle carbon-13 nuclei (Figure 4) were subtracted from each other ($\sigma_3 - \sigma_2$) and dCSs were considered to be $\Delta\delta_{iso} \approx -\Delta\sigma_{iso}$ (see Appendix A). For the NMR property calculations, GIAOs were utilized for treating the gauge dependence problem and threshold-parameters were set to a level that all the orbitals are likely to be included, even linearly dependent ones. The SCF energy threshold, the parameter that controls when the SCF energy is considered to have converged, was changed from the default value 10^{-5} to 10^{-3}. This was done in order to make calculations converge more easily. From previous work the numerical sensitivity to this parameter (SCF energy threshold) was found to be extremely low, around 0.05% for the corresponding change [14]. Therefore, the assumption here is that results are not changed to any significant degree.

3.4 Degree-of-freedom optimization

The size of the training set (600) is a limiting factor for the QCPS-model dimensions. Based on prior experience, this allows for fitting of around 30 variables. These variables are assigned to correspond to the DOFs of the molecule: atom-atom (here, only carbon-carbon) distances (2 atoms), atomic angles (3 atoms) and dihedral angles (4 atoms). The optimal DOF selection is done heuristically (see Figure 3 step 3) by dividing the molecule into symmetry regions (Figure 4). Here, a DOF-goup refers to a small group of DOFs in relation to the active spin pair in order to find out the relevant DOFs for the model performance. Because of the high degree of symmetry in the molecule, the symmetry around the middle ^{13}C-nuclei is emphasized in the selection of the DOFs. Therefore, the smallest group investigated is that of 4 DOFs (see Table 1). The relative importance of a given region is evaluated by computing correlation between the test set and model outputs. DOF-groups having the highest dCS correlation are considered to have the best predictive capability and are considered as most probable candidates for the full 30-DOF model. Further optimizations are done by trial and error using the correlation with the test set as a metric of accuracy. The optimization step is only done using the larger QC-computed training values (aug-pcSseg-3/aug-pcSseg-1). The lighter QC-computed training set (aug-pcSseg-3/pcSseg-0) is only used with the final model to assess the effect of using computationally less demanding training samples to the model's performance (Figure 3 step 4).

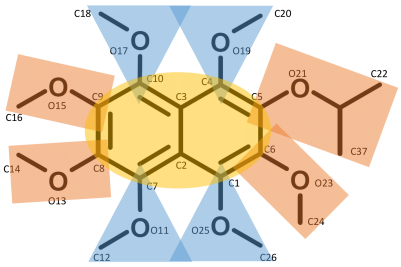


Figure 4: The naphthalene derivative molecule is divided into 3 regions that are based on symmetry around the middle spin pair. The yellow circle indicates the Naphthalene-rings, Blue triangles mark the symmetric side groups and the red rectangles show the non-symmetric side-groups. The distinction non-symmetric/symmetric refers to the fact that in the red rectangle group one of the side groups is different from the rest. Figure also indicates the atom names used in the Z-matrix representation.

3.5 Statistical deviation and the experimental reference

The statistical deviation of results are estimated with Monte Carlo error (see Section 2.10). Both trajectories were divided into five equally long sections and mean was

computed for each section. Then standard deviation of means were calculated and equation (24) was applied.

An experimental measurement of the dCS between the middle carbons with varying temperatures exists [31]. This work compares the slope determined with measurements that have been carried at 14.1 T magnetic field using temperatures 15, 24 and 34 °C, at normal air pressure.

4 Results

4.1 Optimal QCPS model

Table 1 lists the results of the investigation of the most relevant DOFs for the QCPS model performance. It lists all the DOFs used in a given DOF-group (see Section 3.4), the naming convention adopted in this report for the DOF-groups and the model correlation with the test set of QC-computed shielding constants (for C2 and C3) and dCS values. The best dCS correlation for a single DOF-group (0.4840) is achieved with a group that used all the angles in the main rings (RINGang). After that, the second best correlation of 0.2917 comes from using the 5 c-c distances closest to the middle spin pair (5CCnear). The largest correlation coefficients in the table are the C3 and C2 correlation coefficients for the 5CCnear DOF-group (0.6108 and 0.5618, respectively). It appears that the closest c-c distances contain significant information for the individual shielding constants but for some reason, the main ring angles produce larger dCS correlation. Also, DOF-groups containing the DOFs involving the symmetric side groups (closer to the spin pair) appear to produce somewhat larger correlations than the non-symmetric side groups, although all of these correlations are very small. This result is somewhat intuitive if the shielding is assumed to be mainly a local phenomenon [16]. Although the dCS of the spin-pair is a result of the presence of the non-symmetric side groups of the molecule, the geometrical DOFs of those groups do not appear to correlate to that difference directly. Correlations that are negative can also be important. Large negative correlation indicates a negative relationship between the variables [32] so a model can benefit from addition of such DOF-groups. Notable such case here is the RINGdih group, which has a dCS correlation of -0.2050. This DOF-group was included in the best 28-DOF model in this report (Table 2).

Table 2 lists the QCPS models investigated in this report. The models are constructed based on DOF-groups detailed in Table 1. Since the models use parameters from a combination of DOF-groups and have more DOFs in a given model, they are expected to have larger correlations with the test set. The best correlation was achieved with a 28-DOF model E that includes all ring angles and ring dihedral angles, symmetric side group dihedral angles and all the ring c-c distances (see Figure 4). This

selection of DOFs is used on the QCPS model named "Model 2", which is used to determine the $ddCS/dT$-slope. This same group of DOFs also make up the "Model 1", which is optimized with smaller QC-computed training samples. Model 2 produces correlation coefficient of 0.7049 for the dCS. Corresponding correlation with Model 1

Table 1: Naming convention of DOF-groups, descriptions and correlations with a given set of DOF (see Section 3.4). The "DOF" column shows the individual atoms and DOF definitions used in a given DOF-group (see Figure 4), while "Description" describes the DOF-groups verbally. "C2/C3 corr" show how well the QCPS model predictions of CS for carbons 2/3 correlate with the QC test values. "dCS corr" is the corresponding correlation for the dCS. Color coding corresponds to Figure 4. Any given DOF-group contains only DOFs explicitly listed in the "DOF" column

Name	DOF[a]	Description	dCS corr	C2 corr	C3 corr
5CCnear	C2:C1 C3:C2 C4:C3 C7:C2 C10:C3	5 c-c distances closest to the center	0.2917	0.5618	0.6108
4CCfar	C5:C4 C6:C1 C8:C7 C9:C8	Rest of the c-c distances in main rings			
RINGang	C3:C2:C1 C4:C3:C2 C5:C4:C3 C6:C1:C2 C7:C2:C1 C8:C7:C2 C9:C8:C7 C10:C3:C2	All the angles in the main rings	0.4840	0.3317	0.2029
RINGdih	C4:C3:C2:C1 C5:C4:C3:C2 C6:C1:C2:C3 C7:C2:C1:C6 C8:C7:C2:C1 C9:C8:C7:C2 C10:C3:C2:C1	All the dihedral angles in the main rings	-0.2050	0.0944	-0.0539
AangCOC	C14:O13:C8 C16:O15:C9 C22:O21:C5 C24:O23:C6	Non-symmetric side group COC angles	-0.0756	0.1167	0.0177
AangOCC	O13:C8:C7 O15:C9:C8 O21:C5:C4 O23:C6:C1	Non-symmetric side group OCC angles	-0.0756	0.1167	0.0177
AdihCOCC	C14:O13:C8:C7 C16:O15:C9:C8 C22:O21:C5:C4 C24:O23:C6:C1	Non-symmetric side group COCC dihedral angles	-0.0578	-0.2227	0.1571
AdihOCCC	O13:C8:C7:C2 O15:C9:C8:C7 O21:C5:C4:C3 O23:C6:C1:C2	Non-symmetric side group OCCC dihedral angles	-0.0851	0.0208	0.0286
AdistOC	O13:C8 O15:C9 O21:C5 O23:C6	Non-symmetric side group OC distances	0.1834	0.1911	-0.1512
SangCOC	C12:O11:C7 C18:O17:C10 C20:O19:C4 C26:O25:C1	Symmetric side group COC angles	-0.0160	-0.0733	-0.0649
SangOCC	O11:C7:C2 C17:C10:C3 O19:C4:C3 O25:C1:C2	Symmetric side group OCC angles	-0.0181	0.0989	0.1146
SdihCOCC	C12:O11:C7:C2 C18:O17:C10:C3 C20:O19:C4:C3 C26:O25:C1:C2	Symmetric side groups COCC dihedral angles	0.2350	-0.0776	0.1878
SdihOCCC	O11:C7:C2:C1 O17:C10:C3:C2 O19:C4:C3:C2 O25:C1:C2:C3	Symmetric side groups OCCC dihedral angles	0.0365	0.0457	0.0177
SdistOC	O11:C7 O17:C10 O19:C4 O25:C1	Symmetric side groups OC distances	-0.0110	-0.2348	0.1309

a) Two atoms specify a bond length, three atoms an atomic angle and four atoms a dihedral angle

is 0.7050. The correlations of the individual carbons C2 and C3 agree with the model 2 correlations to the 2nd decimal place and are therefore not detailed here. This correlation is relatively poor and suggests that the model has some predictive power but the dimension is not large enough to account for the subtle effect of chemical shift between the middle spins. For model A, the dCS correlation 0.5430 is larger than the correlations for the individual carbons of 0.4473 and 0.2879 for C2 and C3 respectively. However, when the DOF-group 5CCnear is added, the situation is always reversed. This corresponds to what was seen when the individual DOF-groups were compared in Table 1, where the C2 and C3 correlations were largest for this DOF-group.

Table 2: Correlations (for carbon 2, 3 and dCS) for 5 QCPS model considered in this work. "# DOFs" indicates the number of DOFs used in each model, "description" describes the DOF-groups, from which the model is constructed and "DOF-groups" lists the DOF-groups according to the naming convention of Table 1. The color coding corresponds to Figure 4

Name	# DOFs	dCS corr	C2 corr	C3 corr	description	DOF-groups
A	12	0.5430	0.4473	0.2879	Ring angles Symmetric COCC dihedral angles	RINGang SdihCOCC
B	17	0.6656	0.7366	0.7229	Ring angles Symmetric COCC dihedral angles 5 Middle CC distances	RINGang SdihCOCC 5CCnear
C	21	0.5846	0.6922	0.6969	Ring angles Symmetric side group dihedral angles 5 Middle CC distances Non-symmetric COCC dihedral angles	RINGang SdihCOCC 5CCnear AdihCOCC
D	21	0.7029	0.7460	0.7430	Ring angles Symmetric COCC dihedral angles 5 Middle CC distances 4 Outer ring CC distances	RINGang SdihCOCC 5CCnear 4CCfar
E	28	0.7049	0.7468	0.7629	Ring angles Symmetric COCC dihedral angles 5 Middle CC distances 4 Outer ring CC distances Ring dihedral angles	RINGang SdihCOCC 5CCnear 4CCfar RINGdih

4.2 Temperature dependence of chemical shift difference

Figure 5 shows the $\mathrm{d}dCS/\mathrm{d}T$-slopes. Error bars are estimated using Monte Carlo error (see Section 2.10). "Model 1" describes the model build from training samples computed with the smaller QC computations, using aug-pcSseg-3 for the middle spins and pcSseg-0 for the rest of the molecule (aug-pcSseg-3/pcSseg-0). "Model 2" is output of the model built from larger QC calculations (aug-pcSseg-3/aug-pcSseg-1). "Experimental" refers to the experimentally determined slope mentioned in Section 3.5.

Computed mean dCS values for 35°C trajectories agree within the error estimates for both models. It is evident that using reduced QC computations (Model 1 vs. Model 2) change the ensemble average to some extent. When the model (optimized for 35°C trajectory) is used for the 25°C trajectory, the error estimates increase significantly.

Also the fact that error bars do not overlap indicates that the levels of errors are underestimated by the Monte Carlo error. This is natural since the Monte Carlo error only estimates the error using the standard deviations of the MD trajectory means and can therefore only estimate the statistical deviation of the data. The less accurate QC training set computations are one factor that certainly can lower the accuracy of prediction. The error produced by inaccuracies in the training set do not necessarily produce lower precision in the model predictions but certainly can effect the model accuracy. Both models correlate to same extent with their individual test sets but the ensemble averages differ quite a bit. This is evidence for the need to use accurate QC computations for the training samples if accuracy of prediction is required. This is to be expected because the dCS is very subtle for this molecule.

The error ranges in the ensemble averages makes determination of a meaningful slope problematic. Because the error is most likely underestimated, even the sign of the slope can't be determined. The absolute level of dCS is also dependent on the used exchange-correlation functional [14]. This thesis is mainly focused on the slope so this level of accuracy is not a concern at this time.

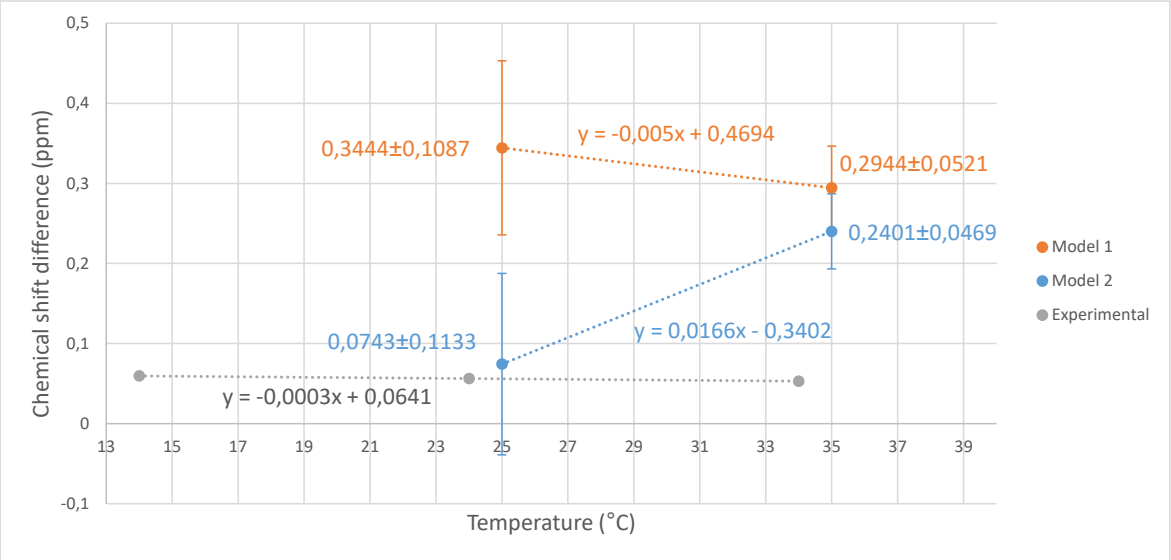
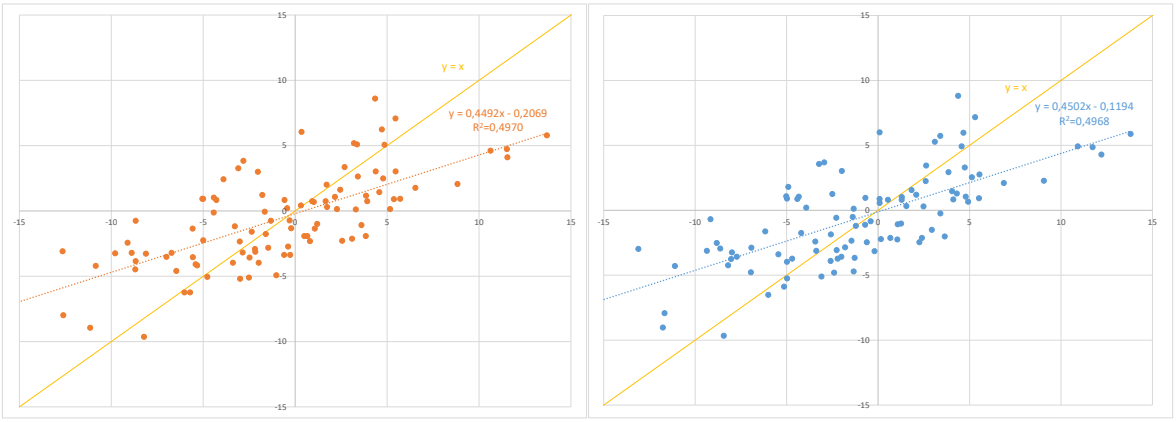


Figure 5: Mean dCS as a function of temperature. Circle-markers indicate the mean dCS values for a given QCPS model and MD trajectory. Error bars are estimated with Monte Carlo error (see Section 3.5). Experimental setup is also described in the same section

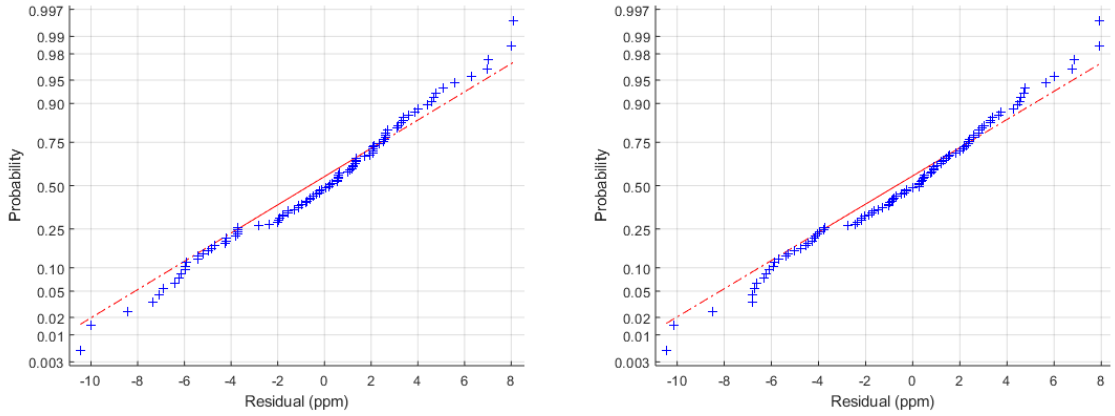
Figure 6 shows the scatter plots of dCS predictions of both models against the their individual QC computed test samples. The reader is reminded that both models contain the same samples but the level of QC computations are different. The predictive capability of both models appears to be very similar. Both models predict dCS values in a smaller range than the QC-computed samples. This is further evidence that the model dimension is not large enough to account for the full ranges of dCS.



(a) Model 1 **(b)** Model 2

Figure 6: Scatter plots of Models 1 and 2 with QC computed test set vs. model predictions. The range of values is smaller for the model predictions than it is for the QC-computed samples. Both models appear to produce very similar predictions for the test set samples.

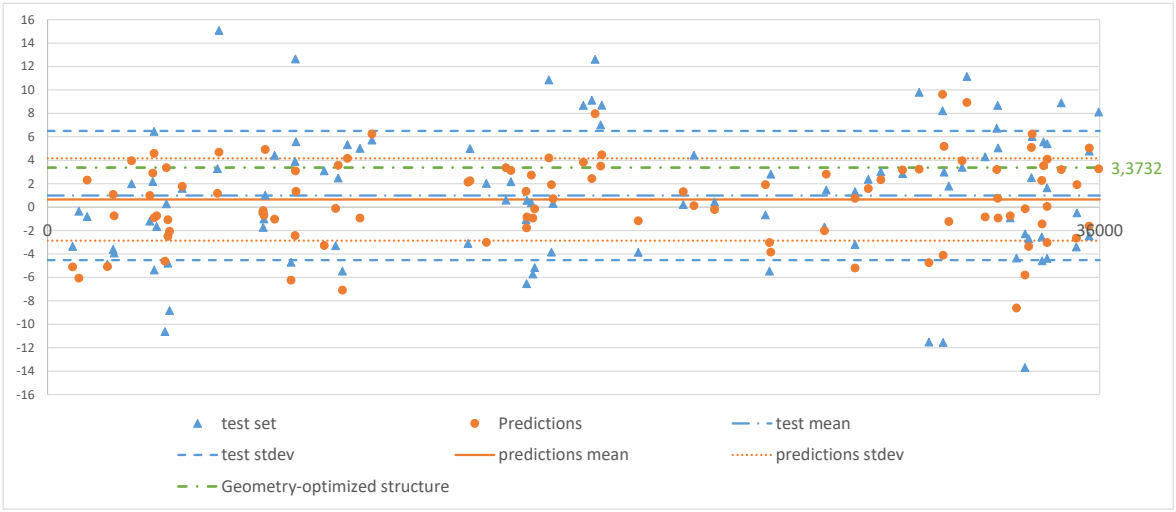
Further analysis that can be done by comparing the model predictions with the test samples is to use residual analysis [32]. Residuals are calculated by subtracting the predicted dCS value from the corresponding QC computed test sample. This produces a vector of residuals that can be evaluated with the matlab function normplot() [33]. The resulting normal probability plot can be used to evaluate if the residuals are normally distributed. Figure 7 shows the normal probability plots of both models. The normal distribution of residuals indicates that there is no systematic under- or overestimation of the dCS predictions. The line in the figure indicates the pure normal distribution of data. By and large, residuals from both models appear normally distributed but the slight deviation in the middle part of the figure might indicate some abnormalities in the data. Deviation of the tail ends are common and do not necessarily indicate abnormalities. Individual points at either ends of plot can indicate outliers [32].



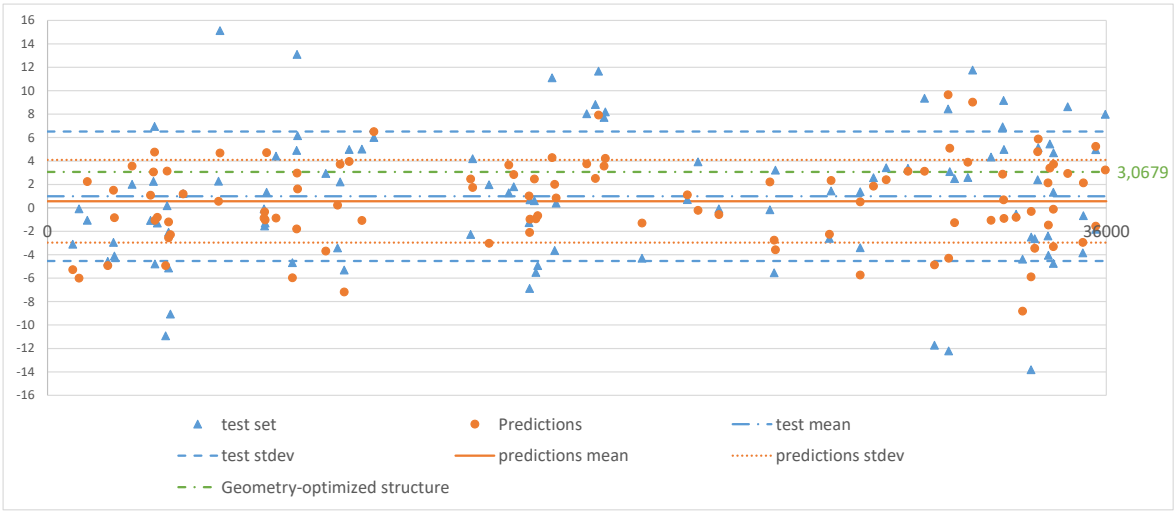
(a) Model 1 **(b)** Model 2

Figure 7: Normal probability plots for both QCPS model output residuals. Residuals are calculated by subtracting the predicted dCS value from the corresponding QC-computed test sample. Data points follow the red line well, indicating that the residuals are normally distributed to a high degree. This suggest that the predictions do not have large systematic prediction errors.

Figure 8 shows the how the test set dCS-values are distributed over the whole MD trajectory. Again, the behavior of both models is similar. Standard deviations of the test set and predictions again show that the test set has larger spread of values. The means of both sets appear to be relatively close, showing that the predictions are not biased. Figure also indicates the QC-computed dCS-values of the geometry-optimized structure (see section 3.2). The dCS value for the geometry-optimized structure is 3.3732 ppm when computed using aug-pcSseg-3/pcSseg-0 basis set and 3.0679 ppm when computed using aug-pcSseg-3/aug-pcSseg-1. These values appear to be quite a bit larger than the predicted mean values.



(a) Training set 1



(b) Training set 2

Figure 8: Subfigures a-b) show the spread of the dCS values of the QC computed test data and model output predictions. Horizontal axis is the sample number in the MD trajectory and vertical axis is the dCS-value. Solid red lines show the mean value of predictions, blue dash-dotted-line shows the mean for the QC-computed test set values. Standard deviations are indicated with the red dotted lines for the QCPS outputs and blue dashed-lines for the QC outputs. Individual datapoints are indicated with blue triangles for the QC test points and red diamonds for the QCPS predictions. dCSs of the geometry-optimized structure, computed with corresponding basis sets, are shown with a green dash-dotted lines and indicated numerically next to the lines. While the means for predictions and test samples are very similar, the standard deviation of the test values is larger. This indicates that the QCPS model is not capable of reproducing the variety of the data and tends to underestimate the absolute magnitudes of the dCS.

In order to evaluate the similarity of QC-computed dCS values, Figure 9 shows the normal probability plot of residuals of both sets of QC-computed values form all 700 configurations. Residuals are computed by subtracting the Set 1 (aug-pcSseg-3/pcSseg-0) sample from the corresponding Set 2 (aug-pcSseg-3/aug-pcSseg-1) sample. Altough

mostly linear, the fact that both tail ends of the figure slope below the line might indicate a slight left skew in the residuals. This could mean that Set 1 overestimates the dCS values to some extent [32]. This observation is supported by the fact that the predicted mean dCS values are larger for Model 1 than for Model 2 (Figure 5).

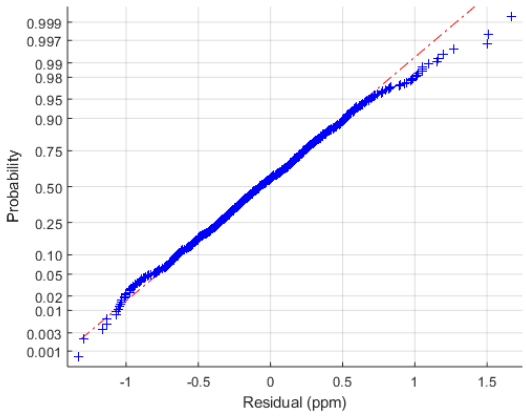


Figure 9: Normal probability curve showing the residuals (Set 1 sample subtracted from the corresponding Set 2 sample) between Set 2 and Set 1 of all 700 QC-computed tensor components. It appears that residuals have little bit of left skew, indicating that the dCS values might be overestimated in Set 1 to some extent.

The possible slight systematic overestimation is also seen in Figure 10 that shows the evolution of mean dCS as a function of number of included samples. The difference in means between the two models is bigger for the 25°C trajectory. This is to be expected because the model is optimized for the 35°C trajectory and should therefore give better predictions for it. The effect of the systematic overestimation for the smaller QC samples is seen in a slight convergence in the 35°C trajectory plots, where the estimates of the mean dCS values appear to separate slightly when the number of samples in the computed mean is increased. The Model 1 predicted mean is larger than the Model 2 prediction for both temperatures. This suggest that accurate QC samples are needed to produce reliable ensemble average.

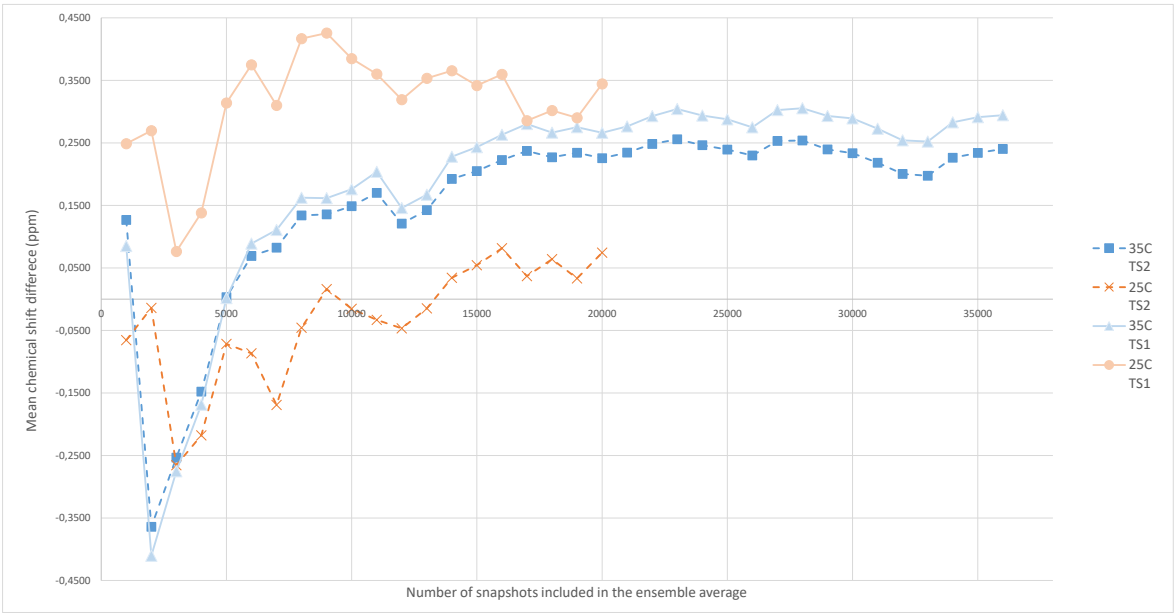


Figure 10: The change in the mean dCS values for both models and temperatures, when the number of samples included in the mean is increased in the increments of 1000 samples. The blue lines (triangles and squares) show the evolution of mean dCS of 35°C trajectories predictions with models 1 and 2 while the red lines (circles and crosses) show the same for 25°C trajectory outputs. The shorter-length 25°C trajectory appears to have more instability for the added samples in the mean dCS values. This indicates that the trajectory should be longer. Also the difference between the model 1 and 2 outputs are greater for the 25°C trajectory. The model 1 systematically predicts higher dCS values than model 2.

5 Discussion

Two 28-DOF QCPS models were constructed and used to efficiently analyze a large number of MD simulation snapshots. Models were trained with QC-computed shielding tensor values, using 1) previously optimized LDBS approach for accuracy vs. computational cost [14] (aug-pcSseg-3/aug-pcSseg-1, Set 2) and 2) a reduced QC computation approach with smaller number of basis functions (aug-pcSseg3/pcSseg-0, Set 1). The two QCPS models were optimized for the 35°C trajectory and used to estimate dCS ensemble averages of 25°C and 35°C trajectories. This yielded a $\mathrm{d}dCS/\mathrm{d}T$-slope that was compared to the experimentally determined slope (Figure 5).

Resulting slopes were dominated by statistical error. This was the case for the Model 1 slope to such an extent that even the sign of the slope could not be accurately determined. Model 2 yielded a slope that has opposite sign as compared to experimental result.

From Figure 5 it is apparent that the statistical deviations are much larger for the 25°C dCS estimates that they are for the 35°C dCS estimates. This is most likely due to the fact that models were optimized for 35°C trajectories only. This problem could

possibly be alleviated by computing a separate training set for the other trajectory and optimizing a separate model for it. Also the fact that the 25°C MD trajectory was smaller than the 35°C one (20000 vs. 36000) could have contributed to these results. Based on Figure 10, evolution of dCS ensemble average with increasing number of included samples, a longer simulation length should be considered in order to minimize the statistical errors from both trajectories. Other possible source of error in the dCS ensemble average predictions could be the selection of training configurations. This thesis used a very simple method to select the samples, a maximization of the Euclidean distances between the samples. another approach might be to use this method as a starting point and then use a parameter sensitivity analysis method such as SOBOL [34] to get a more representative selection of training samples.

This work also studied the different DOF selections for the QCPS model. Because of the computational QC bottleneck, the size of the model dimension is limited by the training set size, making the selection of DOFs for the model important. The DOF analysis suggests that the most important DOFs for the model are the ones located closest to the middle spin pair (Tables 1 - 2). This is somewhat intuitive as shielding is assumed to be mostly locally active phenomenon [16]. The optimal 28-DOF model is still lacking predictive range based on the fact that the spread of the predicted dCS data was smaller than the the spread of corresponding QC test calculations (Figures 6 - 8). This is also reflected by the relatively poor dCS correlation of 0.7049 with the test set (Table 2).

The correlation could possibly be improved by increasing the model dimensions. This would increase the required number of QC computations for the training set. Another possible improvement could be the addition of solvent DOFs to the model. This would require that solvent molecules are also added to the QC computations, increasing the computational demands even further. This would possibly necessitate new investigation on the strategies to reduce computational cost. The LDBS approach used in this work could possibly be used as a starting point for the investigation [14]. One possible approach could be to lean on the side of computational load minimization at the cost of accuracy. This would allow for more training points and thus larger possible model dimensions.

When the models based on different QC-level training sets are compared with residual analysis, a possible slight systematic overestimation may be observed (Figure 9). Although this is by no means a clear indication of skew in the residuals, this observation is supported by the fact that Figures 5 and 10 both show the models optimized for Set 1 to predict larger values for dCS. However, the scatter plots (Figure 6) and the residual analysis with the test set (Figure 7) both show that the model predictions compared to their individual test samples perform almost identically. This is a somewhat intuitive result since the model assumes the training samples to be absolutely accurate. Still,

the predictive ranges of the models and the test set distributions behave similarly for both training sets. This might allow for a reduction in the level of QC computations in order to increase the sampling frequency of the MD trajectory.

The required model dimensions could possibly be investigated by computing lighter QC training set samples. This approach is likely to effect the predicted ensemble average (loss of accuracy) but if the QC computed samples are precise enough (having no wild swings in the results), the correlation with the test set as a metric should allow for a more detailed DOF analysis and an understanding of how big a model is actually needed.

Appendices

Previously presented in author's previous work [14].

A Chemical shift and shielding constant

The chemical shift δ_i of nucleus i expressed with resonance frequencies ν_i is [16]

$$\delta_i = \frac{\nu_i - \nu_{ref}}{\nu_{ref}}. \tag{26}$$

This can also be expressed using shielding constants which are more relevant for this report

$$\delta_i = \frac{\sigma_{ref} - \sigma_i}{1 - \sigma_{ref}}. \tag{27}$$

The shielding constant of nucleus i in the reference standard molecule σ_{ref}, is expressed in ppm units and is very small compared to 1. Therefore, it can be approximated as

$$\delta_i \approx \sigma_{ref} - \sigma_i, \tag{28}$$

where i stands for the nucleus of interest and ref stands for the reference compound [35]. In this case, since the interest is in the chemical shift difference between the two middle carbons, the reference compound cancels out and the relationship with chemical shift difference $\Delta\delta$ and shielding constant difference $\Delta\sigma$ can be regarded as

$$\Delta\delta = -\Delta\sigma. \tag{29}$$

References

[1] T. Maly, G. T. Debelouchina, V. S. Bajaj, et al. "Dynamic nuclear polarization at high magnetic fields". *Journal of Chemical Physics* 128.5 (2008), p. 052211.

[2] S. Nelson, D Vigneron, J Kurhanewicz, et al. "DNP-hyperpolarized ^{13}C magnetic resonance metabolic imaging for cancer applications". *Applied Magnetic Resonance* 34.3 (2008), pp. 533–544.

[3] W. S. Warren, E. Jenista, R. T. Branca, et al. "Increasing hyperpolarized spin lifetimes through true singlet eigenstates". *Science* 323.5922 (2009), pp. 1711–1714.

[4] J. Keeler. *Understanding NMR spectroscopy*. 2nd ed. Chichester: Wiley, 2011, pp. 260–262. ISBN: 1-119-96493-8.

[5] G. Pileio, M. Carravetta, and M. H. Levitt. "Storage of nuclear magnetization as long-lived singlet order in low magnetic field". *Proceedings of the National Academy of Sciences* 107.40 (2010), pp. 17135–17139.

[6] P. Vasos, A Comment, R Sarkar, et al. "Long-lived states to sustain hyperpolarized magnetization". *Proceedings of the National Academy of Sciences* 106.44 (2009), pp. 18469–18473.

[7] M. Carravetta, O. G. Johannessen, and M. H. Levitt. "Beyond the T_1 Limit: Singlet Nuclear Spin States in Low Magnetic Fields". *Physical Review Letters* 92.15 (2004), p. 153003.

[8] G. Stevanato, J. T. Hill-Cousins, P. Håkansson, et al. "A nuclear singlet lifetime of more than one hour in room-temperature solution". *Angewandte Chemie International Edition* 54.12 (2015), pp. 3740–3743.

[9] G. Pileio, J. T. Hill-Cousins, S. Mitchell, et al. "Long-lived nuclear singlet order in near-equivalent ^{13}C spin pairs". *Journal of the American Chemical Society* 134.42 (2012), pp. 17494–17497.

[10] D. Case, I. Ben-Shalom, S. Brozell, et al. "AMBER 2018". University of California, San Francisco. 2018.

[11] P. Håkansson. "Prediction of low-field nuclear singlet lifetimes with molecular dynamics and quantum-chemical property surface". *Physical Chemistry Chemical Physics* 19.16 (2017), pp. 10237–10254.

[12] F. Musil, M. J. Willatt, M. A. Langovoy, et al. "Fast and accurate uncertainty estimation in chemical machine learning". *Journal of Chemical Theory and Computation* 15.2 (2019), pp. 906–915.

[13] K. Nagashima and S. S. Velan. "Understanding the singlet and triplet states in magnetic resonance". *Concepts in Magnetic Resonance Part A* 42.5 (2013), pp. 165–181.

[14] J. Havisto. "Use of locally dense basis sets for chemical shift calculations in naphtalene derivative: Study on the convergence to the CBS limit". Department of Physics, B.Sc. thesis. Oulu. 2019.

[15] D. Chesnut and K. Moore. "Locally dense basis sets for chemical shift calculations". *Journal of Computational Chemistry* 10.5 (1989), pp. 648–659.

[16] P. Atkins. *Atkins' physical chemistry*. 10th ed. Oxford: Oxford University Press, 2014, pp. 513–519, 522–523. ISBN: 978-0-19-969740-3.

[17] P. W. Atkins and R. S. Friedman. *Molecular quantum mechanics*. 5th ed. Oxford: Oxford university press, 2011, pp. 258–260, 296–333. ISBN: 978-0-19-954142-3.

[18] T. W. Keal and D. J. Tozer. "The exchange-correlation potential in Kohn–Sham nuclear magnetic resonance shielding calculations". *Journal of Chemical Physics* 119.6 (2003), pp. 3015–3024.

[19] J. J. Kohanoff. *Electronic structure calculations for solids and molecules: theory and computational methods*. Condensed matter physics, nanoscience and mesoscopic physics. Cambridge: Cambridge University Press, 2006, pp. 195–201. ISBN: 0-511-64831-6.

[20] W. J. Hehre. "Ab initio molecular orbital theory". *Accounts of Chemical Research* 9.11 (1976), pp. 399–406.

[21] F. Jensen. "Polarization consistent basis sets: Principles". *The Journal of Chemical Physics* 115.20 (2001), pp. 9113–9125.

[22] F. Jensen. "Basis set convergence of nuclear magnetic shielding constants calculated by density functional methods". *Journal of Chemical Theory and Computation* 4.5 (2008), pp. 719–727.

[23] F. Jensen. "Segmented contracted basis sets optimized for nuclear magnetic shielding". *Journal of Chemical Theory and Computation* 11.1 (2015), pp. 132–138.

[24] F. Jensen. *Introduction to Computational Chemistry*. Chichester: Wiley, 1999, pp. 192–203. ISBN: 0-471-98085-4.

[25] M. Kaupp, V. G. Malkin, et al. *Calculation of NMR and EPR parameters: theory and applications*. John Wiley & Sons, 2006, pp. 85–100.

[26] D. R. Jones, C. D. Perttunen, and B. E. Stuckman. "Lipschitzian optimization without the Lipschitz constant". *Journal of Optimization Theory and Applications* 79.1 (1993), pp. 157–181.

[27] M. A. Álvarez, L. Rosasco, and N. D. Lawrence. "Kernels for Vector-Valued Functions: A Review". *Foundations and Trends in Machine Learning* 4.3 (2012), pp. 195–266.

[28] A. Feiguin. *Monte Carlo error analysis*. URL: `https://web.northeastern.edu/afeiguin/phys5870/phys5870/node71.html`.

[29] K. Aidas, C. Angeli, K. L. Bak, et al. "The Dalton quantum chemistry program system". *Wiley Interdisciplinary Reviews: Computational Molecular Science* 4.3 (2014), pp. 269–284.

[30] The author wishes to acknowledge CSC – IT Center for Science, Finland, for computational resources.

[31] A. Kantola, measurements made at the NMR Research Unit, University of Oulu.

[32] NIST/SEMATECH. *e-Handbook of Statistical Methods*. URL: `http://www.itl.nist.gov/div898/handbook`.

[33] MATLAB. *9.6.0.1072779 (R2019a)*. The MathWorks Inc., 2019.

[34] A. J. Keane and P. B. Nair. *Computational Approaches for Aerospace Design: The Pursuit of Excellence*. Chichester: Wiley, 2005, pp. 264–265. ISBN: 0-470-85540-1.

[35] M. W. Lodewyk, M. R. Siebert, and D. J. Tantillo. "Computational prediction of ^{1}H and ^{13}C chemical shifts: a useful tool for natural product, mechanistic, and synthetic organic chemistry". *Chemical Reviews* 112.3 (2012), pp. 1839–1862.